DOCUMENT RESUME

ED 385 592 TM 024 040

AUTHOR Livingston, Samuel A.
TITLE An Empirical Tryout of Kernel Equating.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-33
PUB DATE Jul 93
NOTE 40p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equated Scores; High Schools; *High School Students; Sample Size; *Statistical Distributions; United States History
IDENTIFIERS Advanced Placement Examinations (CEEB); *Discrete Variables; Empirical Research; *Kernel Method; Smoothing Methods

ABSTRACT

Kernel equating is a method of equating test scores devised by P. W. Holland and D. T. Thayer (1989). It takes its name from kernel smoothing, a process of smoothing a function by replacing each discrete value with a frequency distribution. It can be used when scores on two forms of a test are to be equated directly or when they are to be equated through a common anchor. The discrete score distributions are replaced with continuous distributions, and then equating is done with the continuous distributions. This "continuization" is accomplished by replacing the frequency at each discrete score value with a continuous frequency distribution centered at that value. The distribution that replaces the discrete function is called the "kernel." Data for the examination of the procedure were taken from responses of 93,283 high school students to multiple-choice questions on the United States History Advanced Placement Examination using samples of 25, 50, 100, and 200 test takers with 50 replications for each sample size. Results support the further study of this approach and the extent to which it can be generalized to other samples. An appendix provides a formula for the root-mean squared deviation. Thirteen figures illustrate the analysis. (Contains 4 references.) (SLD)

RESEARCH REPORT

AN EMPIRICAL TRYOUT OF KERNEL EQUATING

Samuel A. Livingston

Educational Testing Service
Princeton, New Jersey
July 1993

An Empirical Tryout of Kernel Equating

Samuel A. Livingston

ACKNOWLEDGMENTS

I thank Kirsten Yocom for performing the extensive data analysis required by this study and for creating the computer programs necessary to do it efficiently . I also thank Gerald Melican for his help in planning the study, Helen Kahn and Carole Bleistein for their help in obtaining and preparing the data, and Ted Blew for his help in producing the computer-drawn graphs. Support for this research was provided by the Program Research Planning Council at Educational Testing Service.

An Empirical Tryout of Kernel Equating

The purpose of the study

"Kernel equating" is a method of equating test scores, devised by Holland and Thayer (1989). It can be used when scores on two forms of a test are to be equated directly or when they are to be equated through a common anchor. The study described in this report was intended to provide information to answer some practical questions about this method of equating:

How accurate are the equated scores produced by the method?

To what extent does the accuracy of the equated scores depend on the value of a parameter that may be specified by the user?

How accurate are the estimates of the standard errors of the equated scores?

How does the accuracy of the equated scores produced by this method compare with the accuracy of those produced by some other methods?

What is kernel equating?

Kernel equating was devised originally as a solution to a problem arising from the equipercentile definition of equated scores. By this definition, score **x** on Form X and score **y** on Form Y are equated in a population of test-takers if and only if they have the same percentile rank in that population. But in the real world of educational testing, it is rare to find a score on Form Y that has exactly the same percentile rank in the test-taker population as score **x** on Form X.

This problem arises because the score distribution on a given test form is discrete. The problem exists even if the score distributions in the population are known exactly. Holland and Thayer's solution -- kernel equating -- consists of replacing the discrete score distributions with continuous distributions and then equating the continuous distributions. This "continuization" of the distributions is accomplished by replacing the frequency at each discrete score value with a continuous frequency distribution centered at that value. The distribution that replaces the discrete frequency is called the "kernel".[1] Holland and Thayer's kernel equating method uses a kernel that is normal (Gaussian). The continuization of the distributions makes an exact equipercentile equating possible, since it

---

[1]Kernel equating takes its name from kernel smoothing, a general term for the process of smoothing a function by replacing each discrete value with a frequency distribution. Kernel smoothing is sometimes done with a discrete kernel distribution, such as the binomial, in which case it is equivalent to a weighted moving average smoothing.

is always possible to find a score having a specified percentile rank in a continuous distribution.[2]

One desirable feature of Holland and Thayer's kernel equating is that it includes a method for estimating the standard error of the equated score, for any given value of the score to be equated. A description of the method of estimating the standard errors is beyond the scope of this paper. For a presentation of this method, see Holland, King, and Thayer (1989).

The procedure for kernel equating

Holland and Thayer's kernel equating method consists of three essential steps:

1. Estimate the discrete score distributions of the tests to be equated, in the population of test-takers.

2. Replace these discrete distributions with "continuized" distributions, by replacing each discrete frequency with a normal "kernel" distribution.

3. Determine the equipercentile equating relationship between the continuized distributions.

The first step varies, depending on the equating design. In an equivalent-groups design, the score distributions in the population are estimated by smoothing the score distributions observed in the sample, using a log-linear model (Holland and Thayer, 1987). In an anchor equating design, the process of estimating population distributions consists of two sub-steps. The first sub-step is to apply log-linear smoothing to each of the two observed joint distributions (of the score to be equated and the anchor score). The second sub-step is to use the anchor score as a conditioning variable to estimate the population distribution of the scores to be equated. This procedure is sometimes called "frequency estimation". It is based on the assumption that the conditional distribution of each score to be equated, conditioning on the anchor score, is the same in the population as in the smoothed sample distribution.

The second step in this procedure -- the continuization step -- specifies the form of each kernel distribution (normal) and its mean (the discrete score) but leaves its variance unspecified. The variance of all the kernel distributions is controlled by a single parameter, called "h". The value of this parameter can have a great influence on the shape of the continuized distribution. If the value of h is small, e.g., 0.3, the individual kernel distributions will be quite narrow, with little overlap, and the continuized distribution that results from combining them will be spiky in shape. If the value of h is larger, e.g., 1.0, the kernel distributions will

---

[2]The common procedure of using linear interpolation to determine the equated scores in equipercentile equating is mathematically equivalent to continuizing the distribution of scores on the reference form (the "old form") with kernel distributions that are uniform and non-overlapping.

overlap considerably, and the continuized distribution will be quite smooth (extending beyond the range of scores actually observed). As h gets larger, the individual kernel distributions overlap even more, and the continuized distribution tends to resemble a normal distribution.[3] Since the equipercentile relationship between any two normal distributions is linear, an equating of two continuized distributions produced with large values of h will tend to be approximately linear. Holland and Thayer have suggested a criterion for choosing a value of h to continuize a given discrete distribution: choose the value that minimizes the sum of squared differences between the discrete frequencies and the corresponding densities of the continuized distribution (Holland and Thayer, 1989, pp. 30-33). This criterion is programmed into the computer programs that were used in the present study.

The procedure for the study

The method and the data for this study were identical to those for an earlier study (Livingston, 1993) on the accuracy of another method of equating in an anchor design. The study was designed to create a situation in which the equating relationship in the population was known. The data were taken from the responses of 93,283 high school students to the multiple-choice section of the Advanced Placement Examination in United States History. From the 100 items in this section of the examination, the investigators constructed two overlapping subforms of 58 items each. The overlap consisted of 24 items appearing in both subforms to serve as an anchor for equating. The subforms were constructed to be as similar in content as possible, while differing systematically in difficulty. The more difficult of the two subforms was labeled "Form A"; the less difficult subform was labeled "Form B". From each test taker's responses to the items, the investigators computed three scores: a score on Form A, a score on Form B, and a score on the 24-item anchor test.

The distributions of scores on the two subforms in the full population of test-takers indicated a substantial difference in difficulty. The mean scores were 29.2 items correct (50%) on Form A and 35.9 (62%) on Form B. The standard deviations were 8.7 items on Form A and 8.8 on Form B. Thus, the difference in the mean scores was about three-fourths of a standard deviation. The population distribution of scores on Form A showed a slight positive skew; the distribution of scores on Form B showed a substantial negative skew.

The next step was to determine the equating relationship between Forms A and B in the test-taker population. This step was accomplished by a direct equipercentile equating of the observed score distributions on Forms A and B in the entire group of 93,283 test-takers, with no continuization or smoothing. The anchor test played no part in this criterion equating. The

---

[3]Although Holland and Thayer's kernel equating involves no assumption about the form of the distribution of scores on either form of the test, a user who chooses a large value of h is, in effect, assuming that the scores have approximately normal distributions in the population.

results of this direct, full-population equating served as the criterion for evaluating the results of the equatings based on samples of examinees.[4]

The rest of the study consisted of selecting samples of examinees, applying Holland and Thayer's kernel equating method in the samples, and comparing the kernel equating results with the results of the criterion equating. The procedure was as follows:

1. Select two samples of test-takers, by simple random sampling without replacement. Arbitrarily associate each sample with one of the two subforms. For test-takers in the "Form A" sample, treat the score on Form B as unknown; for test-takers in the "Form B" sample, treat the score on Form A as unknown.

2. Use the data from these two samples to estimate the population distributions of scores on the tests to be equated, by performing the smoothing and frequency estimation steps described above. In the smoothing step, preserve the first bivariate moment (the correlation of the test score and the anchor score) and the first three univariate moments of each variable (mean, standard deviation, and skewness). If the sample size is at least 100, also preserve the fourth univariate moment (kurtosis) of each variable.[5]

3. Continuize the estimated population distributions, using h-values computed to minimize the squared-difference criterion. Use these distributions to equate Form A to Form B. Also estimate the standard errors of the equated scores by Holland and Thayer's procedure.

4. Repeat Step 3, using the arbitrarily specified h-value of 0.3 for both distributions.

5. Repeat Step 3, using the arbitrarily specified h-value of 1.5 for both distributions.

6. Replace the test-takers into the population available for sampling.

The study involved fifty replications of this procedure under each of four sample-size conditions: samples of 200, 100, 50, and 25 test-takers. These small sample sizes reflect the investigators' concern about the accuracy of equating in small samples of examinees and their hope that Holland and Thayer's method might offer an improvement over the methods now used.

---

[4]With responses from more than 90,000 test-takers available, the investigators decided that an equating of unsmoothed distributions would provide the best available criterion. In retrospect, it appears that an equating of smoothed distributions would have provided a better criterion, if the smoothing model closely preserved the overall shape of the distributions.

[5]These choices of the number of moments to preserve were based on the results of the earlier study (Livingston, 1993).

The results

To evaluate the accuracy of the equating in a study such as this one, it is necessary to compare the results of the anchor equatings in the samples of test-takers with the results of the direct equating in the full population. This comparison requires a statistic that summarizes the differences between the sample results and the population result. A statistic that is commonly used to summarize such differences is the root-mean-squared deviation (RMSD). Because the accuracy of the equating tends to differ greatly from one part of the score range to another, the RMSD was computed separately at each score level on the form to be equated (Form A). For clarity, this statistic will be referred to as the "conditional RMSD". The formula appears in the appendix to this report. The smaller the conditional RMSD at a particular score level, the more accurate the equating at that score level. The values of the conditional RMSD tend to trace a smooth curve, lower in the middle of the score distribution, where data are plentiful, and higher at the ends, where data are sparse.

Figure 1 shows twelve such curves. For each of four sample sizes (200, 100, 50, and 25 test-takers) there are three curves, for the three different levels of the h-parameter. The units of the conditional RMSD are raw-score points on Form B, the reference form; one point equals approximately 0.1 standard deviations. The percentiles indicated on the horizontal axis refer to the distribution of raw scores on Form A in the full population. The graphs in Figure 1 show a clear effect of sample size on the accuracy of the equating, as might be expected, but very little effect for the choice of h-values. (The computed values of the h-parameter ranged 0.60 to 0.74, as compared with the specified values of 0.3 and 1.5). There was a tendency for the equatings based on an h-value of 1.5 to be slightly less accurate at the upper end of the distribution and slightly more accurate at the lower end of the distribution than the equatings based on smaller h-values. Otherwise, the choice of h had essentially no effect on the accuracy of the equating.

Figures 2 to 5 show the bias in the equatings, computed under the assumption that an unbiased equating procedure would tend to replicate the direct equipercentile equating in the full population. Under this assumption, the bias at each score level is indicated by the difference between the population equating result and the mean of the fifty sample equating results. The curves in Figures 2 to 5 show some irregularities at the low and high ends of the score scale -- below the 1st percentile and above the 99th percentile of the population distribution -- and the irregularities are similar for all four sample size conditions. The source of the irregularities appears to have been the population equating used as a criterion, which did not involve pre-smoothing of the score distributions.

In Figures 2 to 5, the value of the h-parameter does appear to have a noticeable effect -- one that is easy to explain. The equating in the population was strongly curvilinear. A linear equating would have produced equated scores that were higher at the ends of the score range and slightly lower in the middle. In kernel equating, the larger the h-value, the more similar the continuized distributions, and, therefore, the more nearly linear the equating transformation. For any pair of samples of test-takers, the

kernel equatings using h = 1.5 would tend to be more nearly linear than the equatings using the smaller h-values. Therefore, the equatings using h = 1.5 would tend to produce equated scores that were higher at the ends of the score scale and slightly lower in the middle than the equated scores based on the smaller h-values. That is exactly what Figures 2 to 5 show. In some cases the larger h-value resulted in a smaller bias, in some cases a larger bias, and in one case, an approximately equal but opposite bias.

Holland and Thayer's kernel equating method produces not only an equated score for each raw score on the form to be equated, but also an estimate of the standard error associated with it. Under normal circumstances, the accuracy of these estimates would be impossible to determine, but the special conditions of this study provide another way to estimate the standard error of the equated scores at each score level. The fifty replications of the equating (each with different samples of test-takers) under each set of conditions (sample size, h-value) provide a distribution of fifty equated scores on Form B for each score on Form A. The standard deviation of this distribution is an empirical estimate of the standard error of equating at that score level.

The fifty replications also provide a distribution of fifty separate Holland-Thayer estimates of the standard error of equating. Figures 5 to 9 compare this distribution with the single empirical estimate at each score level. Instead of attempting to show the entire distribution of fifty Holland-Thayer estimates at each score level, the figures show selected percentiles of the distribution. Figures 5 to 9 show these results only for the kernel equatings using the computed h-values; the results for the specified h-values of 1.5 and 0.3 were similar. The empirical estimates and the Holland-Thayer estimates of the standard error tended to agree fairly well in the middle of the score range, where there was plenty of data. However, the Holland-Thayer estimates in the smaller samples (50 or less) tended to be systematically larger than the empirical estimate in the upper half of the score range, with the difference becoming substantial above score 45 (the 95th percentile of the population distribution).

Figures 10 to 13 compare the accuracy of the kernel equating results with the results of two other equating procedures, applied to the same data in an earlier study (Livingston, 1993). Both were "chained" procedures, in which Form A was equated to the common-item anchor test in one sample of test-takers, the anchor test was equated to Form B in the other sample of test-takers, and the composition of the two resulting functions was taken as the function equating Form A to Form B. The equatings were equipercentile equatings, using linear interpolation to determine the equated scores. One procedure was an equating of the observed distributions. The other was an equating of smoothed distributions, using the same smoothing procedure as in the kernel equatings in the present study. The kernel equating results shown in Figures 10 to 13 are those produced by the computed h-values, but, as shown previously in Figure 1, either of the two specified h-values (1.5 or 0.3) would have produced nearly identical results.

Figures 10 to 13 show that the equating of smoothed distributions -- with or without the continuization step -- produced much more accurate results

than the equating of observed (unsmoothed) distributions. The full kernel equating procedure provided some additional accuracy at the low end of the score range (below the 5th percentile of the population distribution), when applied to the larger samples of 100 or more test-takers.

The methods represented in Figures 10 to 13 include a chained equipercentile equating of smoothed discrete distributions and a kernel equating of distributions produced by frequency estimation (conditioning on the anchor score), followed by a continuization step. If the difference in the accuracy of these two methods were larger, it would be interesting to determine how much of the difference was attributable to the continuization procedure and how much to the frequency estimation approach vs. the chained approach. However, as Figures 10 to 13 show, in the portions of the score range where most of the data are found, the difference in accuracy is quite small. A more fruitful direction for further research would be to investigate the extent to which the results of this study can be generalized to other tests, to other populations of test takers, and especially to situations in which the groups of test-takers taking the two forms to be equated differ systematically in ability.

References

Holland, P. W. and Thayer, D. T. (1987) Notes on the use of log-linear models for fitting discrete probability distributions. Program Statistics Research Technical Report No. 87-79. Princeton, NJ: Educational Testing Service.

Holland, P. W., King, B. F., and Thayer, D. T. (1989) The standard error of equating for the kernel method of equating score distributions. Program Statistics Research Technical Report No. 89-83. Princeton, NJ: Educational Testing Service.

Holland, P. W. and Thayer, D. T. (1989) The kernel method of equating score distributions. Program Statistics Research Technical Report No. 89-84. Princeton, NJ: Educational Testing Service.

Livingston, S. A. (1993) Small-sample equating with log-linear smoothing. Journal of Educational Measurement, 30, 23-39.

Appendix:
Formula for the Root-Mean-Squared Deviation

Let j index the pairs of samples of a given size: j = 1, 2, ... , 50.

Let x represent a score on Form A.

Let y_x represent the score on Form B that equated to x in the direct equating in the population.

Let $\hat{y}_{xj}$ represent the score on Form B that equated to x in the anchor equating in the jth sample.

The conditional RMSD at score x is computed by

$$RMSD(x) = \sqrt{\frac{1}{50}\sum_{j=1}^{50}(\hat{y}_{xj} - y_x)^2} \quad .$$

Figure 1. RMSD of Equated Scores

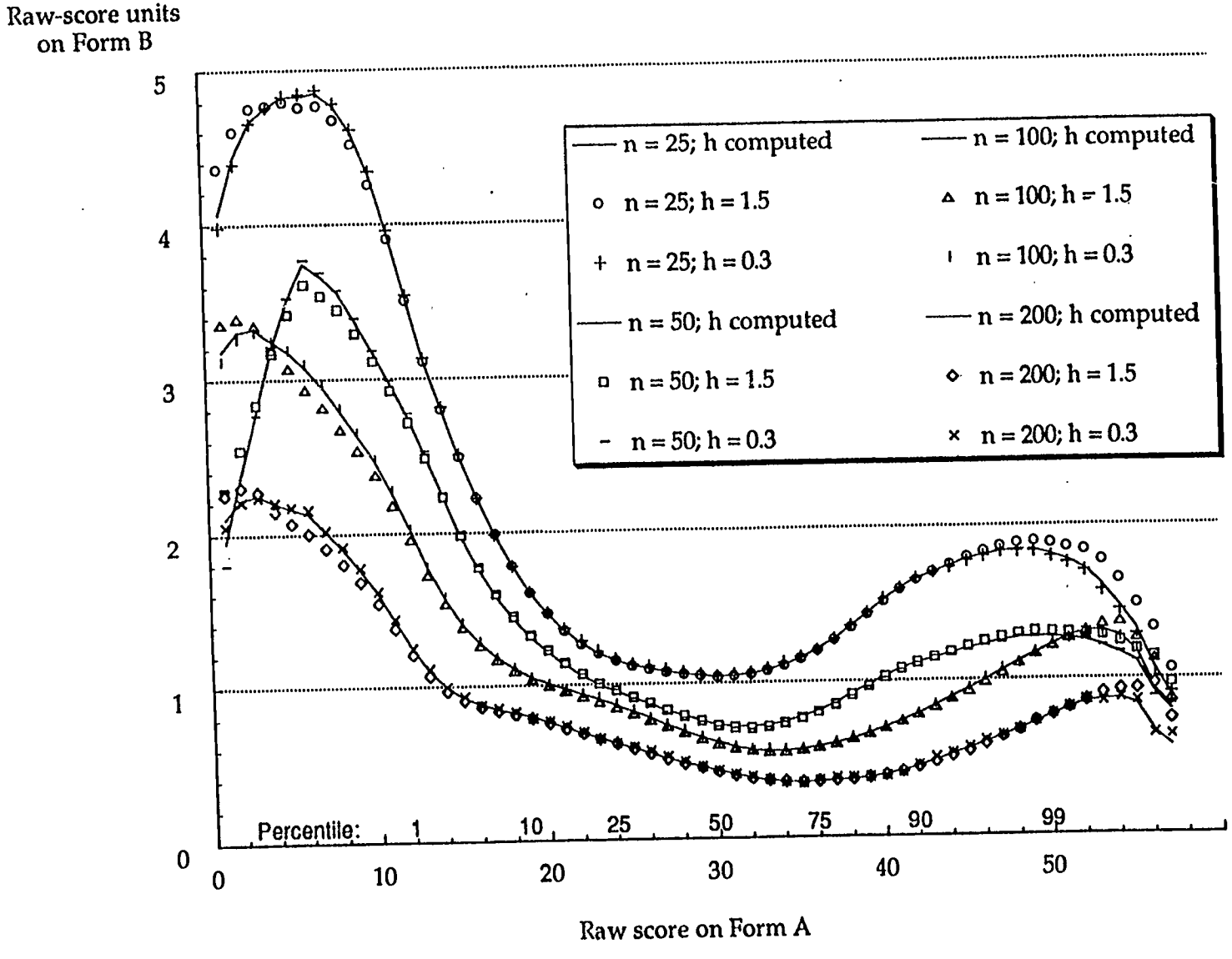

Figure 2. Bias in Equating Results:
Samples of 200 Examinees

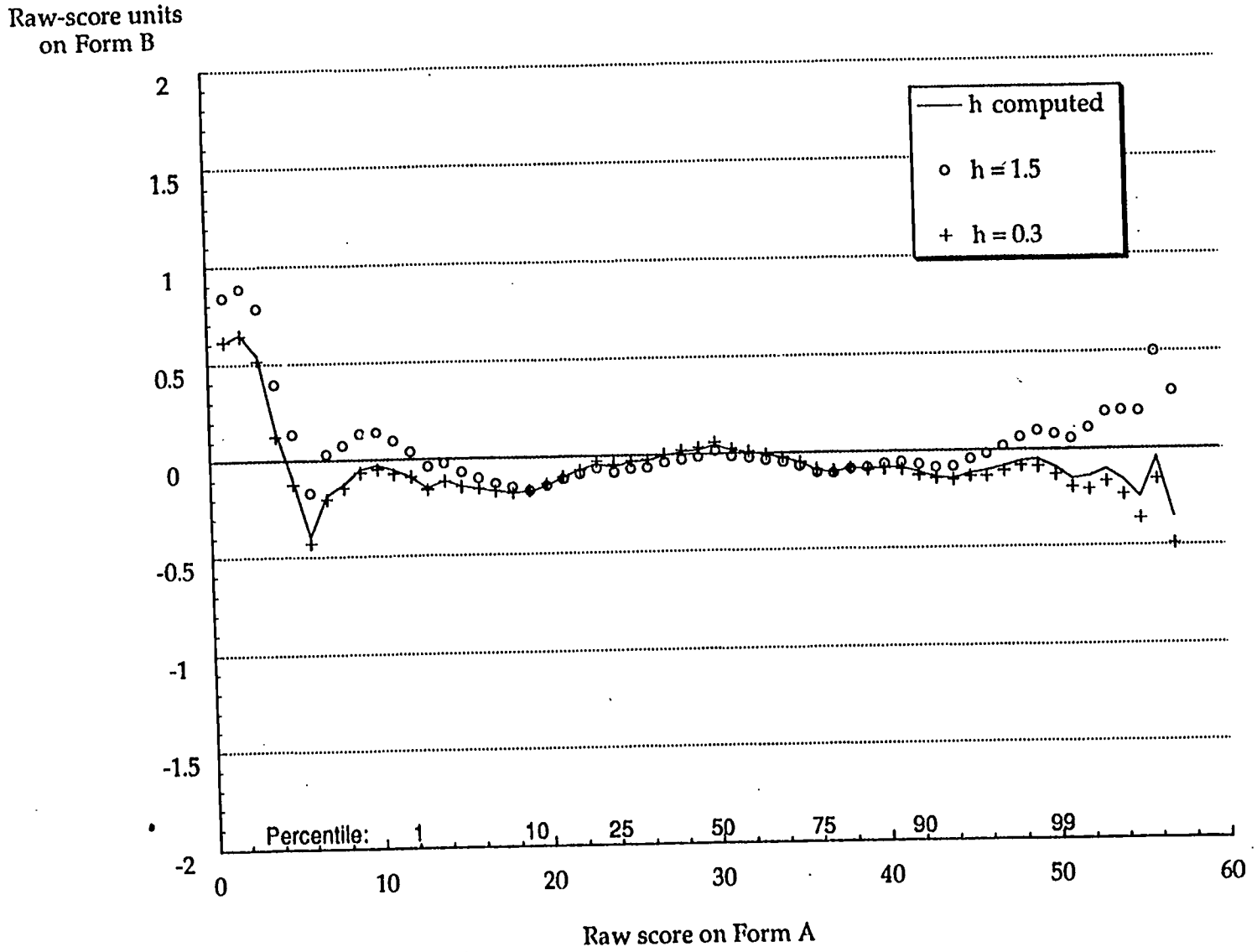

Figure 3. Bias in Equating Results:
Samples of 100 Examinees

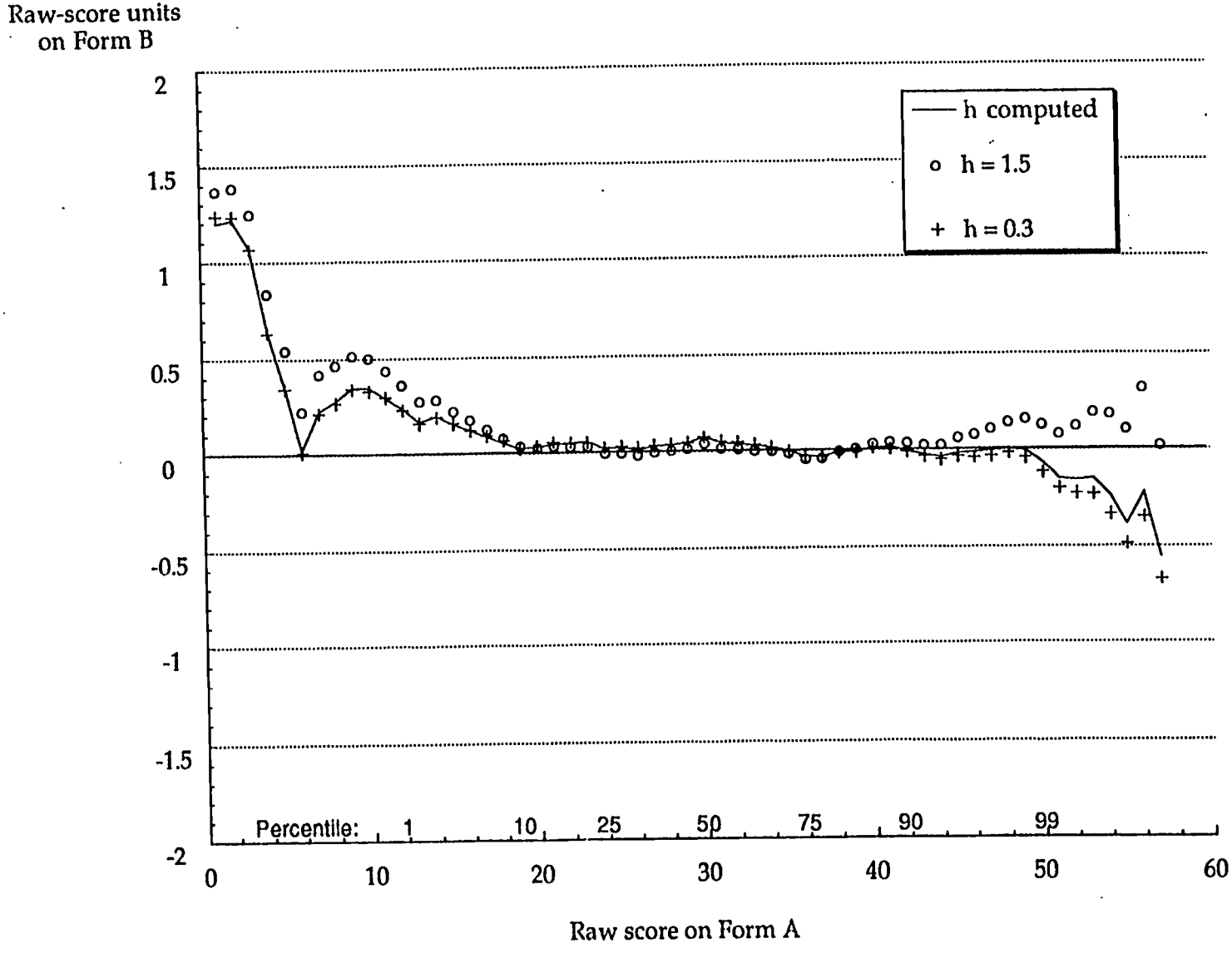

Figure 4. Bias in Equating Results:
Samples of 50 Examinees

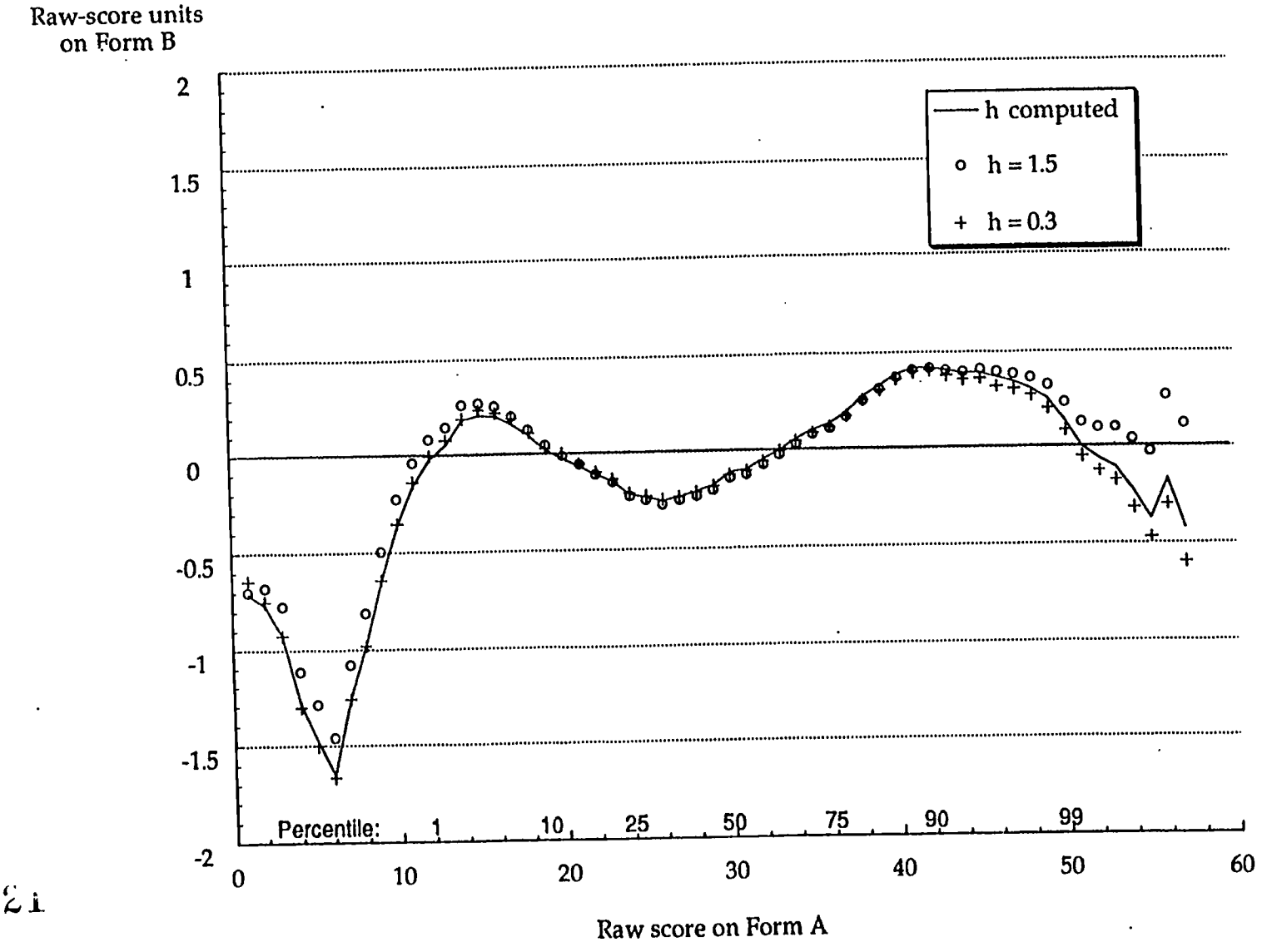

Figure 5. Bias in Equating Results:
Samples of 25 Examinees

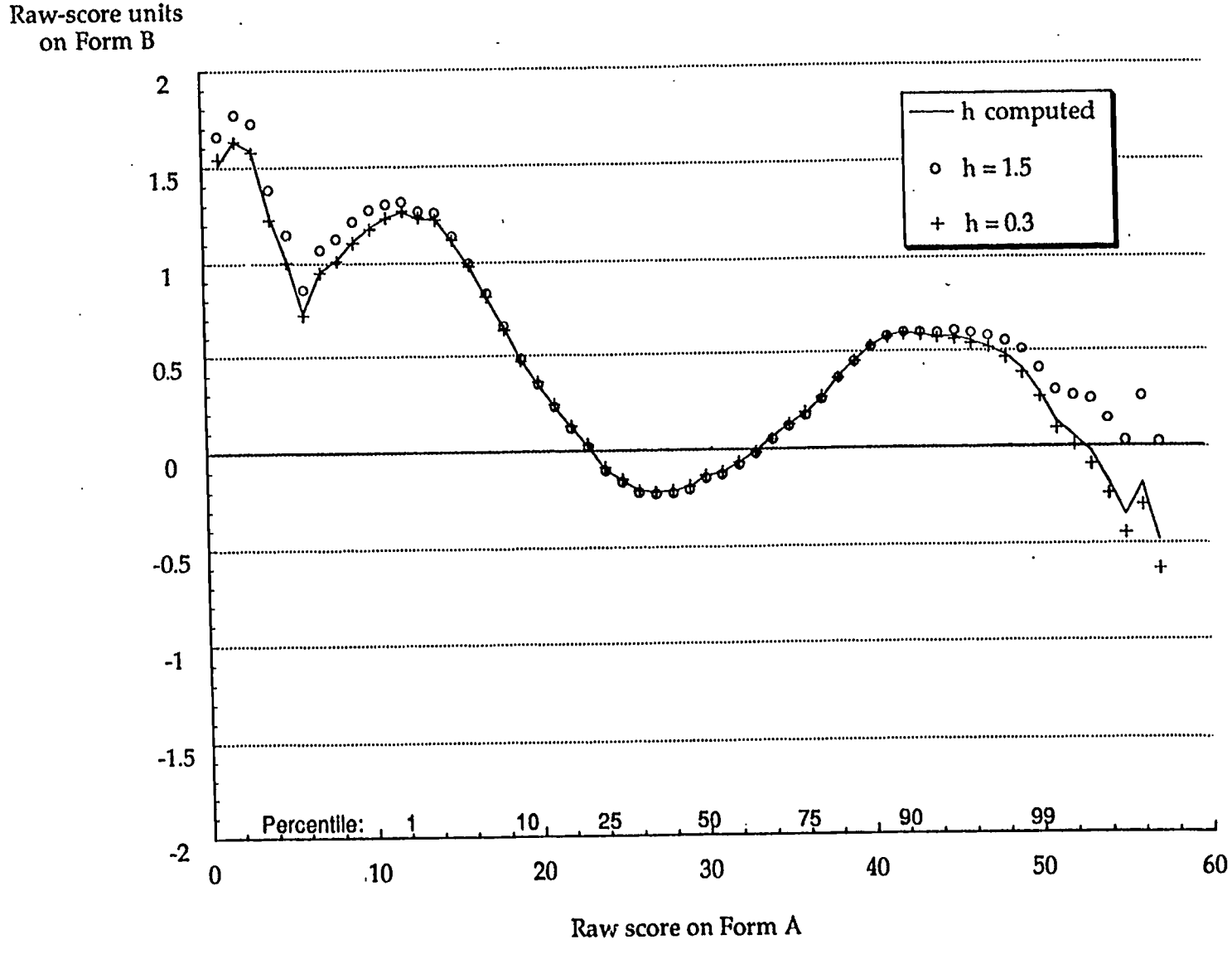

Figure 6. Estimates of the Standard Error of Equating:
Samples of 200 Examinees
h computed

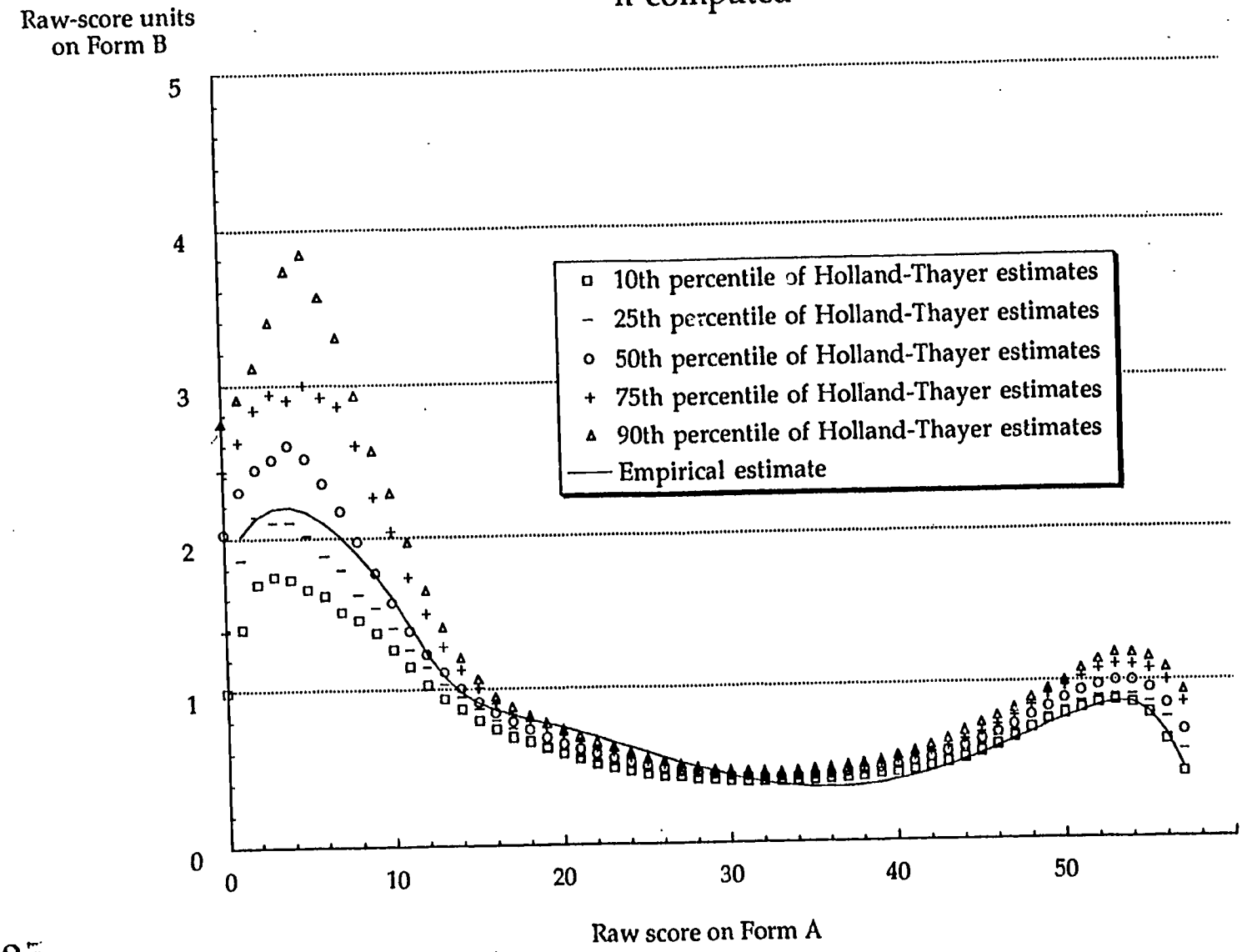

Figure 7. Estimates of the Standard Error of Equating:
Samples of 100 Examinees
h computed

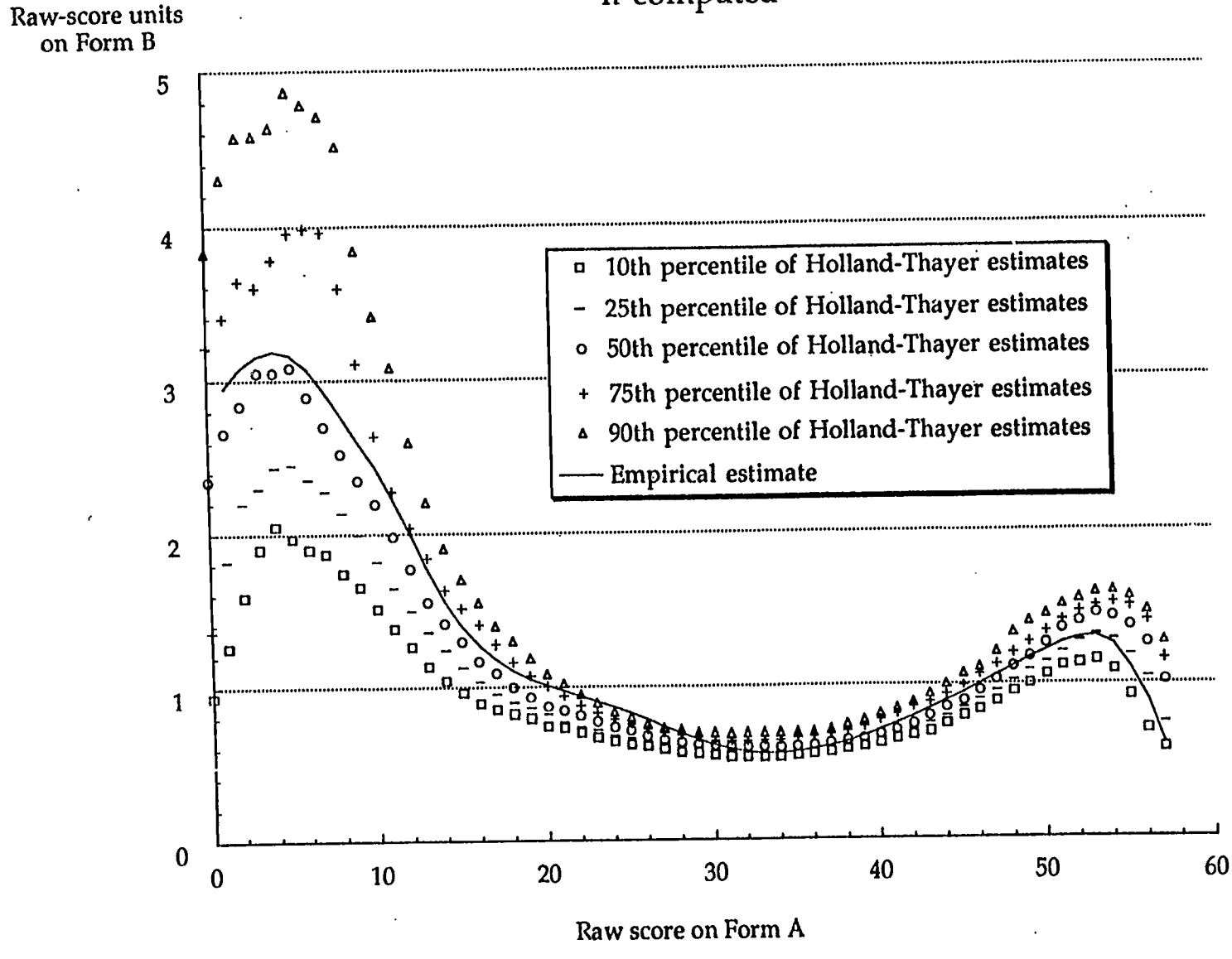

Figure 8. Estimates of the Standard Error of Equating:
Samples of 50 Examinees
h computed

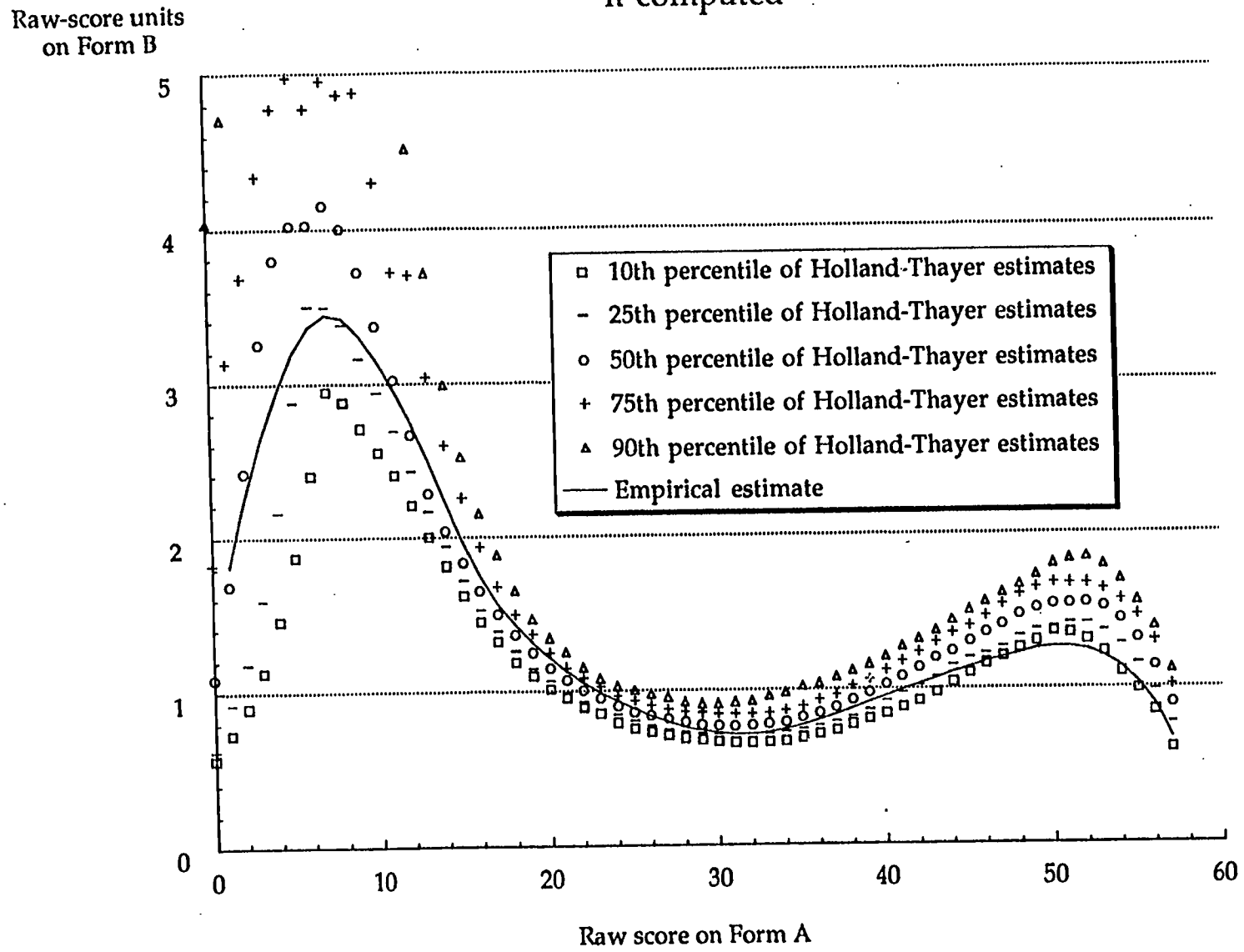

Figure 9. Estimates of the Standard Error of Equating:
Samples of 25 Examinees
h computed

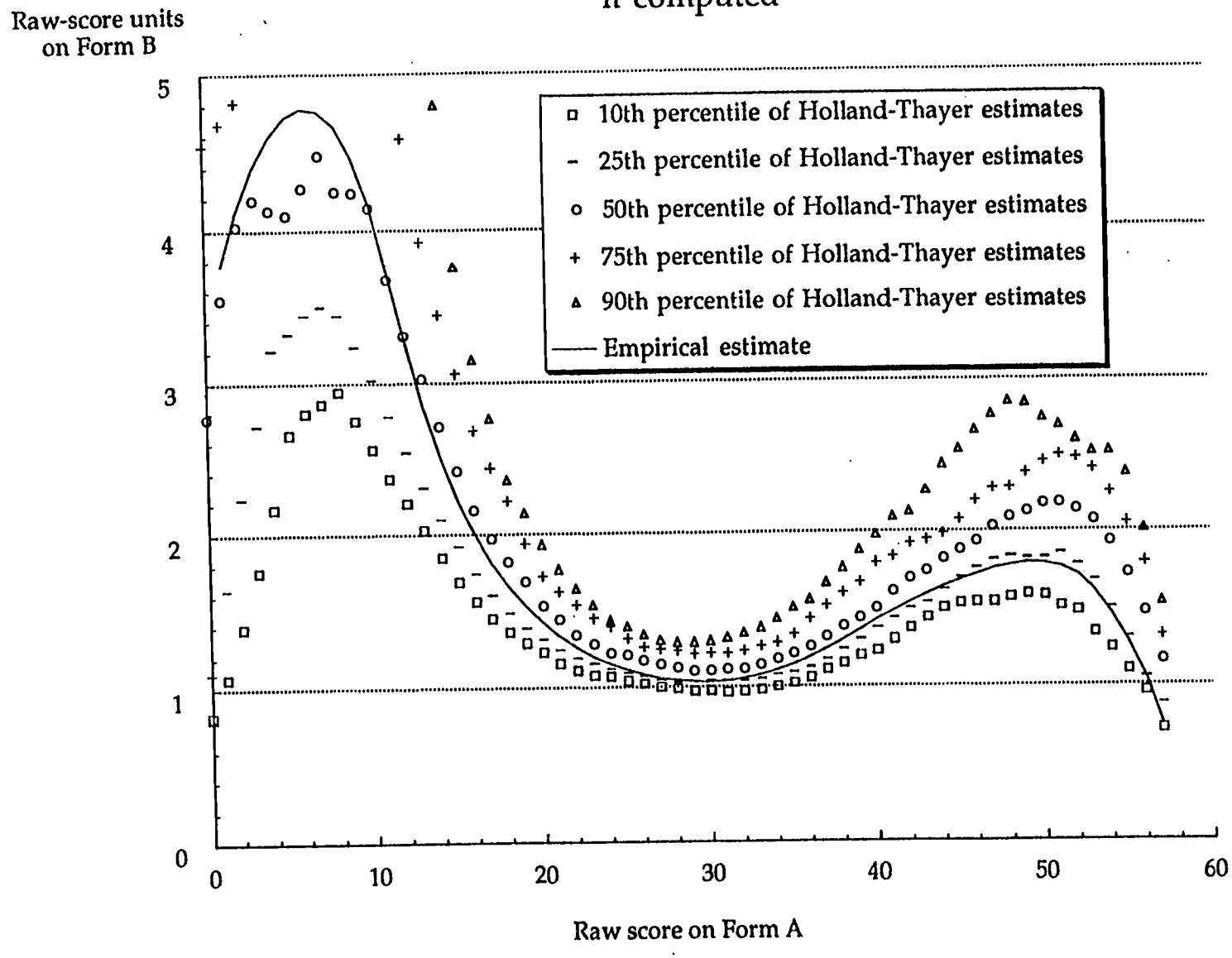

Figure 10. RMSD of Equated Scores:
Samples of 200 Examinees

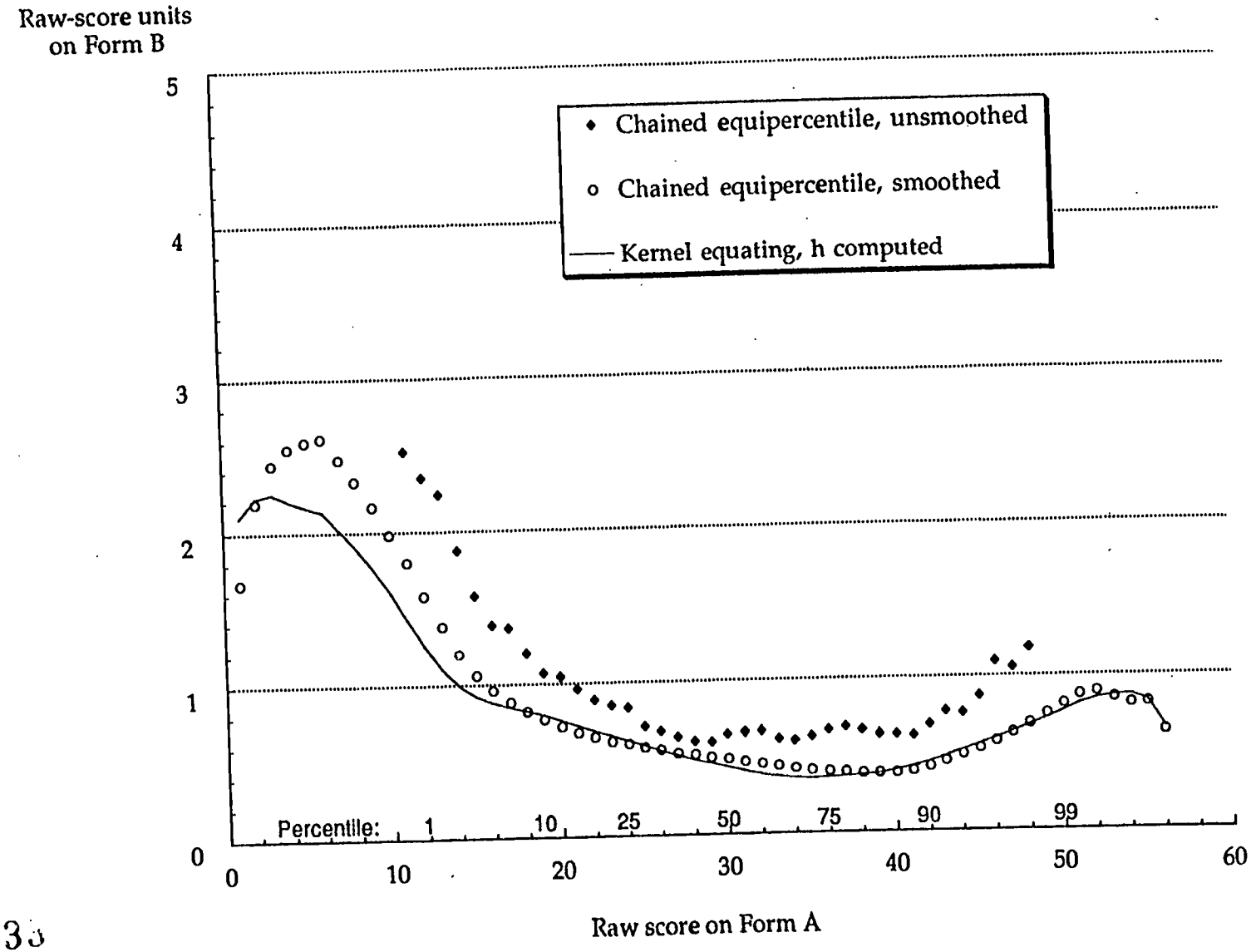

Figure 11. RMSD of Equated Scores:
Samples of 100 Examinees

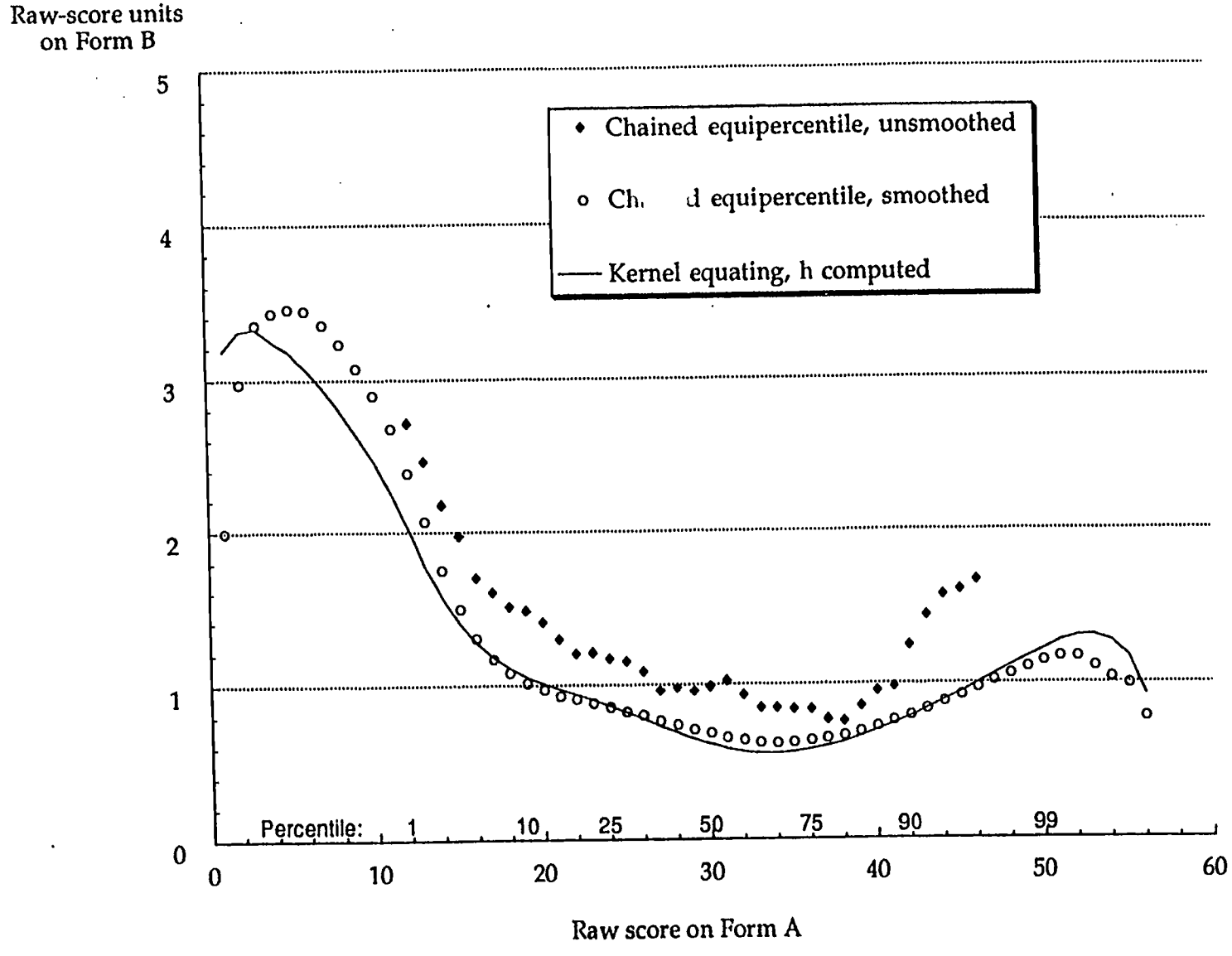

Figure 12. RMSD of Equated Scores:
Samples of 50 Examinees

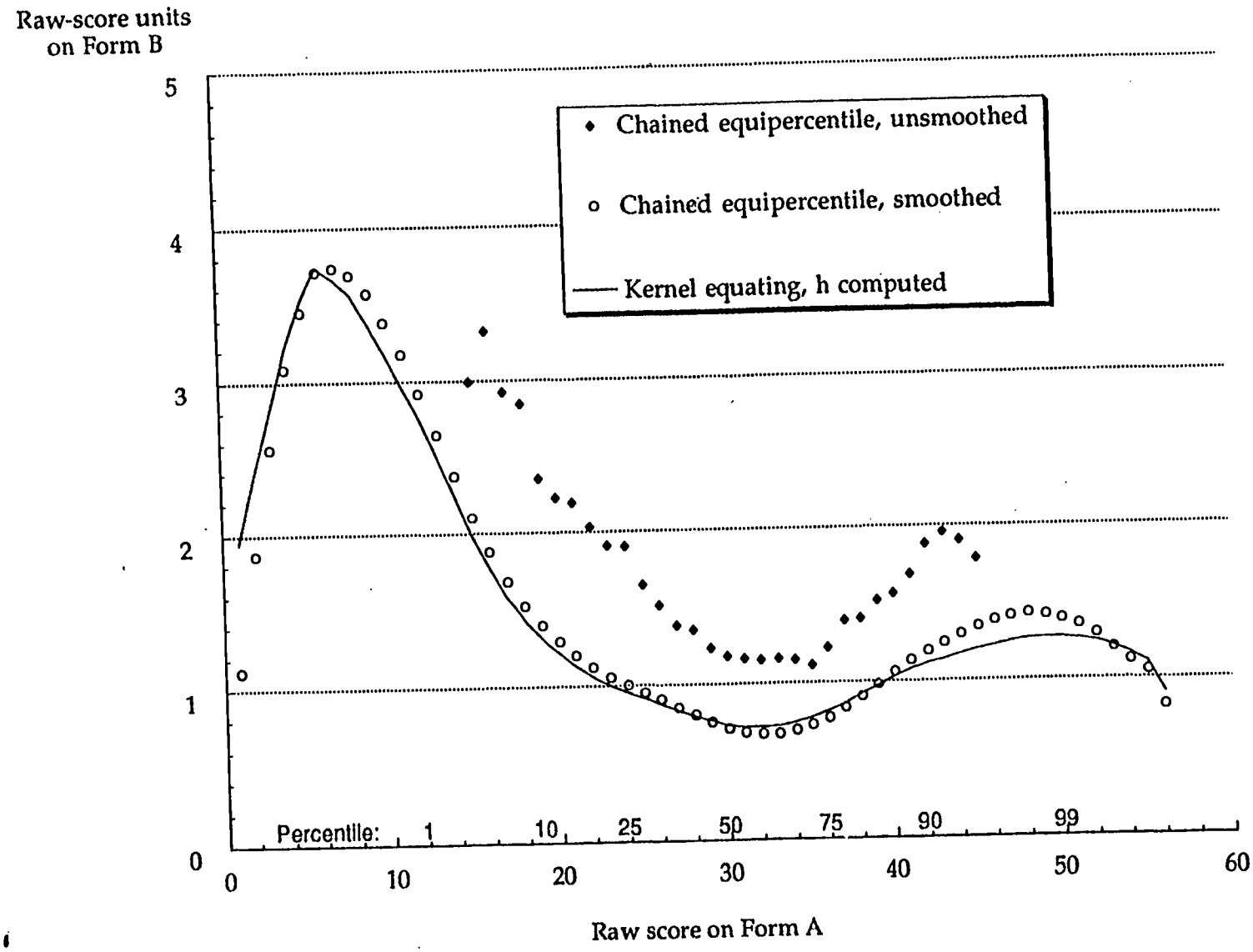

Figure 13. RMSD of Equated Scores:
Samples of 25 Examinees